图书情报工作
第 62 卷 第 18 期　2018 年 9 月

基于加权关键词的领域热点与趋势分析新方法*

■ 奉国和[1]　孔泳欣[2]　肖洁琼[1]

[1] 华南师范大学经济与管理学院信息管理系　广州 510006　[2] 南开大学商学院信息资源管理系　天津 300071

摘要：［**目的/意义**］为克服关键词绝对词频分析的局限性，以关键词多因素加权及得分排名实现领域热点与趋势探索。［**方法/过程**］构建年度－关键词频次矩阵，用水平加权和垂直加权处理关键词词频，设计相对词频模型，计算关键词加权综合分值，以获得更有效的关键词排序。［**结果/结论**］基于关键词加权排序，可以识别量高质优型、量低质优型和突变型关键词，有利于挖掘研究热点和分析趋势。

关键词： 词频分析　加权关键词　热点研究　趋势分析

分类号： G250

DOI：10.13266/j.issn.0252－3116.2018.18.011

关键词是表达文献主题概念的自然语言词汇。一个学术研究领域较长时域内大量学术研究成果的关键词集合，可以揭示研究成果的内容特征，了解学术研究的发展脉络与发展方向[1]。因此，统计关键词在某一类学术文献中所出现的频次，可以判别该学术领域的研究热点，分析发展趋势[2]。词频分析法是基于统计数据，具有客观性、准确性；在一定程度上摆脱定性方法的个人主观性而更具有可信性，因而被广泛地应用于揭示各学科领域的研究热点和发展动态[3]。

随着词频分析法广泛应用于各学科领域，该类文献的数量不断增加，但同时也呈现出较为严重的方法滥用及模板化现象[3]。部分研究仅限于对词频的简单统计和粗略分析，不能通过其数据结果揭示出学科领域知识的内在规律。词频分析法具有广泛的应用性和推广性，但其应用存在一些弊端，因此需要对词频分析法本身进行完善与研究。笔者结合年度总词频数和该关键词总词频数，提出加权关键词模型，以更为准确客观地揭示学科热点和趋势。同时，以我国图书情报学研究领域为例，验证该方法的有效性。

1　相关研究

词频分析法是基于揭示或表达文献核心内容的关键词或主题词在某一研究领域文献中出现的频次高低来确定该领域研究热点和发展动向的文献计量学方法[4]。虽然热点分析类文章常用词频分析法、引文分析法、文献增长率等多种文献计量方法，而应用词频分析法的文献占所有应用了各类文献计量学的热点分析类文献的 61%，是热点分析类文献最常用的文献计量方法[3]。同时，在应用词频分析的文献中，以关键词作为词频分析统计要素的相关文献占有绝对比例，取决于关键词具有直接获得和无需分词的特点[5]。

多数研究成果以关键词的自然频率作为研究的基础和依据，考虑到关键词的非规范化问题，部分研究成果从关键词频次计算、关键词选择、结果分析三个方面改进计量方法以准确揭示词频波动规律。在改进关键词频次计算方面，对于基于绝对词频的统计分析，倪丽娟运用词频绝对值描述研究现状和揭示热点趋势[6]。在基于样本总量变化引起误差的算法改进方面，为消除不同年份论文数波动所造成的影响，邱均平利用篇均关键词频次，以关键词各年出现的频次除以当年的文献总数量来判断其增长或衰减情况[7]。巩永强基于关键词频率探究变化趋势，即某一关键词占当年关键词总数的比例[4]。基于不同样本的数据处理，苍宏宇提出关键词频次标准化处理 Z-Score，以消除国内外文献数相差较大造成的影响[8]。在改进关键词选择方法方面，对于基于研究领域特有特征的热点分析，G. Chen 等结合关键词的人气指数和领域关联度指标来选择关键词[9]。在基于低频项加权方面，E. S. Atlam

＊ 本文系国家社会科学基金项目“基于文本挖掘的科技文献知识发现研究”（项目编号：16BTQ071）研究成果之一。

作者简介：奉国和（ORCID：0000－0002－0774－1544），教授，博士，E-mail：ghfeng@163.com；孔泳欣（ORCID：0000－0003－3653－2415），硕士研究生；肖洁琼（ORCID：0000－0001－5783－2893），硕士研究生。

收稿日期：2018－01－02　修回日期：2018－06－28　本文起止页码：102－109　本文责任编辑：徐健

等提出负权函数和负加权反动词频率函数,提高关键词回收率和精确度[10]。G. Chen 等比较传统术语频率(TF)方法、TF－逆文档频率(TF-IDF)和 TF－关键字活动指数(TF-KAI)这三种方法,得出 TF-KAI 在关键词选择的质量和数量上更为出色[11]。在结果分析改进方面,李珊珊等按照关键词词频划分为低频区、中频区、高频区三个等级,运用定量研究方法定性地分析了文献内在规律及研究热点[12]。上述方法适用不同应用场景,然而目前文献在处理绝对词频时,均立足于年度总词频和年度总论文数对关键词频的影响,并未考虑自身占比情况。自身占比能兼顾削弱词频在数值上的优势,且反映出自身变化率。笔者结合年度总词频数和该关键词总词频数,提出加权关键词模型,探索新的研究思路。

2 加权关键词相对词频模型

关键词年度分布可以反映历年的研究重点,而关键词随时间增长可以反映历年研究热点。笔者将关键词年度分布和关键词每年词频比重有机结合,首先构建年度－关键词的词频矩阵,依据矩阵水平与垂直两个维度加权处理关键词词频,得到相对词频计算公式,以期准确反映关键词的年度分布。然后,确定综合加权关键词排序分数,获得更有效的关键词排序。笔者将该方法称为加权关键词相对词频模型(Weighted Relative Keyword Frequency Model,WRKFM)。

2.1 相对词频计算

构建年度－关键词的词频矩阵,定义函数 $f(i,j)$ 为第 j 年关键词 i 的频次,那么所有年所有关键词的频次可以用矩阵(1)来表示:

$$\begin{bmatrix} f(1,1) & \cdots & f(1,m) \\ \vdots & \vdots & \vdots \\ f(n,1) & \cdots & f(n,m) \end{bmatrix}$$

矩阵(1)

为反映相同年度不同关键词频次强度及不同年度相同关键词频次强度,对关键词频数进行两个维度加权处理:

(1)垂直加权,即关键词当年词频除以当年总词频。设定 n_j 为第 j 年总关键词量,反映到矩阵(1)中,即矩阵(1)第 j 列元素均乘以 $\frac{1}{n_j}$,用矩阵(2)来表示:

$$\begin{bmatrix} f(1,1)\times\frac{f(1,1)}{n_1} & \cdots & f(1,m)\times\frac{f(1,m)}{n_m} \\ \vdots & \vdots & \vdots \\ f(n,1)\times\frac{f(n,1)}{n_1} & \cdots & f(n,m)\times\frac{f(n,m)}{n_m} \end{bmatrix}$$

矩阵(2)

(2)水平加权,计算某关键词当年词频在该关键词统计时间段内总数占比。设定 m_i 为关键词 i 总频次,反映到矩阵(1)中,即矩阵(1)第 i 行元素均乘以 $\frac{1}{m_i}$,用矩阵(3)来表示:

$$\begin{bmatrix} f(1,1)\times\frac{f(1,1)}{m_1} & \cdots & f(1,m)\times\frac{f(1,m)}{m_1} \\ \vdots & \vdots & \vdots \\ f(n,1)\times\frac{f(n,1)}{m_n} & \cdots & f(n,m)\times\frac{f(n,m)}{m_n} \end{bmatrix}$$

矩阵(3)

2.2 加权关键词相对词频模型设计

根据矩阵(2)和矩阵(3),加权关键词相对词频模型可用矩阵(4)来表示:

$$\begin{bmatrix} f(1,1)\times\frac{f(1,1)}{n_1}\times\frac{f(1,1)}{m_1} & \cdots & f(1,m)\times\frac{f(1,m)}{n_m}\times\frac{f(1,m)}{m_1} \\ \vdots & \vdots & \vdots \\ f(n,1)\times\frac{f(n,1)}{n_1}\times\frac{f(n,1)}{m_n} & \cdots & f(n,m)\times\frac{f(n,m)}{n_m}\times\frac{f(n,m)}{m_n} \end{bmatrix}$$

矩阵(4)

为得出更加科学、客观、准确的数据结果,并将其有效转化为知识结论,笔者设计的 *WRKFM* 计算步骤如下:

步骤 1:确定时域,统计关键词及其频次,构建年度－关键词的词频矩阵,计算矩阵(4)的结果;

步骤 2:计算矩阵(4)中每行元素数值之和,即 n 个关键词的相对词频 W,进行由高到低的排序,得出高频关键词;

步骤 3:单独观察矩阵(4)每行的数值,并描绘出其变化趋势,即为关键词的相对词频变化趋势,以预测发展趋势;

步骤 4:根据步骤 2 排序名次,分析其与原绝对词频排序名次之差,有利于监测突变型关键词。

该模型的主要特点具体如下:①相对词频增加了某年度词频比重的影响力,克服单一从样本容量改进词频的不足。因此,若时域内某关键词的总绝对词频高且总体变化大,则其相对词频较大;②相对词频突显某年度对该关键词在该年占比大且绝对频次高的数据,弱化绝对词频低的数据,更易于探测出具有发展潜力的关键词;③低词频的相对词累计频排名变化量与突变主题类型有表征关系,可侧面探测突变词,补充低频词的利用价值。

3　实证分析

利用上述模型，对图情领域文献进行对比分析。在 CNKI 和 CSSCI 上下载 2012 – 2016 五年间的 18 种图情领域核心期刊刊载的文献信息，人工去除无作者、通讯稿、征文稿等非学术类期刊文献，经统计、去重得到 24 618 篇文献。使用 EXCEL 统计，最终得到 34 553 个关键词，人工合并 82 组同义词，去除 120 个无意义词，以下研究选取绝对词频大于等于 5 的关键词(共 2 164个)。

3.1　关键词加权计算

按照矩阵(4)对关键词进行加权计算，列出 2012 – 2016 五年内相对词频值排名前 50 的关键词，如表 1 所示：

表 1　绝对词频与加权相对词频部分结果

排序	关键词	绝对词频						相对词频					
		2016 年	2015 年	2014 年	2013 年	2012 年	累计总和	2016 年	2015 年	2014 年	2013 年	2012 年	累计总和
1	图书馆	294	298	341	420	447	1 800	0.750	0.830	1.107	1.976	2.325	6.988
2	高校图书馆	304	238	272	329	308	1 451	1.029	0.525	0.697	1.178	0.943	4.373
3	公共图书馆	176	190	198	198	188	950	0.305	0.408	0.411	0.392	0.328	1.843
4	大数据	174	109	93	53	9	438	0.639	0.167	0.092	0.016	0.000	0.915
5	数字图书馆	60	72	100	111	190	533	0.022	0.040	0.094	0.123	0.603	0.881
6	信息服务	54	63	85	93	114	409	0.020	0.035	0.075	0.094	0.170	0.395
7	网络舆情	83	64	83	51	43	324	0.094	0.046	0.089	0.020	0.011	0.259
8	图书馆学	36	63	87	57	75	318	0.008	0.044	0.104	0.028	0.062	0.246
9	阅读推广	83	78	52	45	26	284	0.107	0.094	0.025	0.015	0.003	0.245
10	竞争情报	39	46	56	62	98	301	0.010	0.018	0.029	0.038	0.146	0.243
11	知识管理	30	31	44	72	82	259	0.006	0.006	0.017	0.069	0.100	0.197
12	知识服务	43	38	55	77	70	283	0.015	0.011	0.030	0.077	0.057	0.190
13	云计算	18	31	54	59	81	243	0.001	0.007	0.033	0.041	0.102	0.184
14	影响因素	71	42	53	53	68	287	0.066	0.015	0.026	0.025	0.051	0.183
15	微博	42	51	57	74	58	282	0.014	0.027	0.033	0.069	0.032	0.175
16	美国	62	49	69	50	41	271	0.047	0.025	0.061	0.022	0.012	0.166
17	学科服务	44	39	50	66	61	260	0.017	0.013	0.024	0.053	0.041	0.148
18	本体	34	42	43	60	70	249	0.008	0.017	0.016	0.042	0.065	0.147
19	社会网络分析	38	45	61	54	60	258	0.011	0.020	0.044	0.029	0.039	0.144
20	情报学	28	40	54	46	66	234	0.005	0.015	0.034	0.020	0.058	0.132
21	学科馆员	26	21	35	40	74	196	0.005	0.003	0.011	0.016	0.097	0.131
22	知识图谱	52	38	49	54	57	250	0.030	0.012	0.024	0.030	0.035	0.131
23	图书馆服务	53	45	56	49	47	250	0.032	0.021	0.035	0.023	0.019	0.130
24	文献计量	49	21	43	53	48	214	0.029	0.002	0.019	0.033	0.024	0.108
25	知识共享	31	26	43	47	59	206	0.008	0.005	0.019	0.024	0.047	0.103
26	移动图书馆	32	39	52	47	27	197	0.009	0.017	0.036	0.025	0.005	0.092
27	可视化	34	36	51	48	27	196	0.011	0.013	0.034	0.027	0.005	0.090
28	关联数据	48	38	36	42	33	197	0.030	0.016	0.012	0.018	0.009	0.084
29	图书馆联盟	27	22	29	54	45	177	0.006	0.003	0.007	0.043	0.024	0.083
30	大学图书馆	30	37	36	48	33	184	0.008	0.016	0.013	0.029	0.009	0.074
31	信息资源	14	11	22	44	48	139	0.001	0.001	0.004	0.029	0.037	0.072
32	专利分析	28	24	51	34	28	165	0.007	0.005	0.040	0.011	0.006	0.070
33	服务模式	21	20	38	37	47	163	0.003	0.003	0.017	0.015	0.030	0.067
34	信息素养	44	27	31	37	36	175	0.026	0.006	0.009	0.014	0.012	0.067
35	引文分析	28	21	33	42	43	167	0.007	0.003	0.011	0.021	0.022	0.065
36	共词分析	32	20	43	30	34	159	0.011	0.003	0.025	0.008	0.012	0.059
37	实证研究	30	14	30	45	29	148	0.010	0.001	0.009	0.030	0.008	0.057

奉国和，孔泳欣，肖洁琼. 基于加权关键词的领域热点与趋势分析新方法[J]. 图书情报工作,2018,62(18):102－109.

（续表 1）

排序	关键词	绝对词频						相对词频					
		2016 年	2015 年	2014 年	2013 年	2012 年	累计总和	2016 年	2015 年	2014 年	2013 年	2012 年	累计总和
38	信息检索	21	25	34	24	46	150	0.003	0.006	0.013	0.004	0.030	0.057
39	微信	39	28	31	11	0	109	0.029	0.011	0.014	0.001	0	0.055
40	机构知识库	24	31	42	24	26	147	0.005	0.011	0.025	0.005	0.006	0.052
41	数据挖掘	22	30	29	31	40	152	0.004	0.010	0.008	0.009	0.020	0.051
42	图书馆员	20	25	35	32	36	148	0.003	0.006	0.015	0.011	0.015	0.049
43	社会网络	15	18	29	37	37	136	0.001	0.002	0.009	0.018	0.017	0.048
44	企业	25	17	22	31	41	136	0.006	0.002	0.004	0.011	0.024	0.046
45	数字资源	25	23	33	32	34	147	0.006	0.005	0.012	0.011	0.013	0.046
46	科学数据	35	28	28	26	8	125	0.018	0.010	0.009	0.007	0.000	0.044
47	全民阅读	36	31	24	10	20	121	0.020	0.014	0.006	0.000	0.003	0.044
48	电子政务	25	9	22	35	35	126	0.007	0.000	0.004	0.016	0.016	0.043
49	突发事件	38	29	23	17	17	124	0.024	0.011	0.005	0.002	0.002	0.043
50	信息行为	22	28	36	30	23	139	0.004	0.009	0.017	0.009	0.004	0.043

根据表 1，人工筛除边缘性、非核心关键词，相对词频排名突出的包括传统研究方向和研究热点。

传统研究包括“高校图书馆”“公共图书馆”“数字图书馆”“信息素养”“信息检索”“知识管理”。

研究热点具有以下主题：①图书馆服务，包括“信息服务”“阅读推广”“知识服务”“学科服务”“学科馆员”“知识共享”“移动图书馆”“图书馆联盟”等；②情报学工具与应用，包括“云计算”“社会网络分析”“竞争情报”“知识图谱”“文献计量”“可视化”“数据挖掘”等；③信息资源，包括“大数据”“网络舆情”“关联数据”“机构数据库”“社会网络”“电子政务”“突发事件”等。

3.2 加权关键词相对词频变化趋势

相较于绝对词频变化趋势，相对词频的变化趋势视觉效果更佳，对该关键词的某年占比大且绝对频次高的数据更为敏感，倾斜程度更大，同时弱化绝对词频较小的年度词频，以便容易抓取变化率大的关键词，更容易探测出具有发展潜力的关键词。从图 1 中可以看出，“大数据”在 2016 年大幅上升，其绝对词频为 174，约为 5 年总词频的 40%，因此“大数据”是量高质优型关键词，突显出相对词频变化图对量高质优型关键词的有效抓取。“阅读推广”的绝对词频在 2012－2014 年量低、增长率显著，在 2014－2016 年量高、增长缓慢，在图 2 中表现为前端发展平缓、后段发展迅猛，表现出相对词频变化图对关键词的宏观把握。

综合图 1 和图 2 可得：①增长型关键词有“大数据”“阅读推广”“网络舆情”，如“大数据”和“网络舆情”呈现出不同的增长方式，“大数据”是快速增长状态，从 2012 年的 9 次到 2016 年的 174 次，“网络舆情”是缓慢增长状态，此类关键词的增长与现行科研大环境相吻合，时代需求结合紧密；②波动型有“公共图书馆”“高校图书馆”“知识管理”，如“高校图书馆”在 2013 年处于波峰，2015 年处于波底谷，2012 年和 2016 年数量大致持平；③下降型有“竞争情报”“知识服务”“学科服务”“社会网络分析”，如“竞争情报”从 2012 年的 98 次到 2016 年的 39 次，呈明显下滑状态，该主题在本研究设定时间域之前属于热点，但后续研究热度下降。增长型关键词更大概率成为未来研究趋势。图 1 和图 2 中的曲线表明，波动型和下降型占比较大，需要扩大关键词范围，以检索更多增长型关键词，以便更好地预测未来研究趋势。

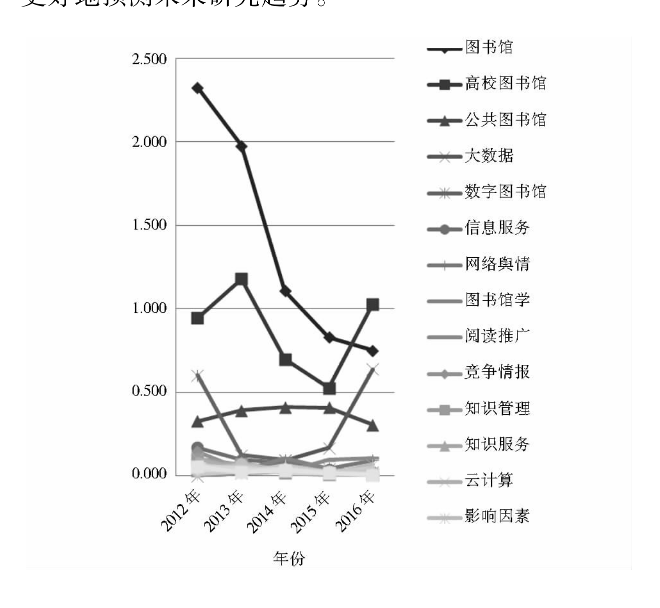


图 1　相对词频变化趋势部分统计

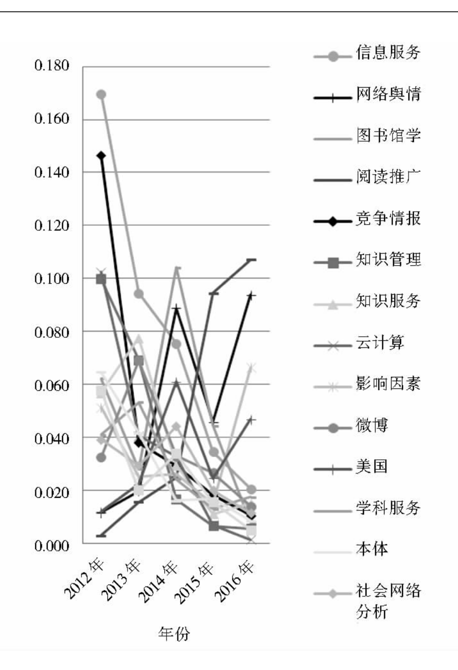


图 2 图 1 下部分放大显示

3.3 WRKFM 与关键词仿真结果对比

3.3.1 高频关键词实验结果 针对相对词频排序前 100 的关键词，对比相对词频的排名和绝对词频的排名，选取两者排名差绝对值前 10 名的关键词得到表 2。其中，表 2 中负数为该关键词在相对词频中排名低于绝对词频排名，正数则相反。

根据表 2，抽取累计绝对词频总和相当的关键词对进行分析（见表 3）：①“数据库”与“情报分析”，“数据库”于 2012 年度数值突出，后 4 年明显递减，而“情报分析”的最大值较小，但总体都处于较高的水平。虽然两者的绝对词频相同，但是“情报分析”的相对词频明显高于“数据库”（见图 3）。②“信息需求”与“社交网络”，虽两者的总量和分布基本相似，但“信息需求”的极端值与平均值相差较大，影响了相对词频的累计总和（见图 4）。③“读者服务”和“微信”，虽然“微信”的 2012 年度值为 0，但其增长趋势较大，突显其发展潜

表 2 高频关键词排名对比情况

关键词	数据库	对策	国家图书馆	信息需求	比较研究	读者服务	图书情报学
排名差	-22	-20	-15	-15	-9	-9	-9
关键词	信息资源	用户行为	移动服务	公共文化服务	期刊评价	微信	
排名差	11	12	15	23	25	26	

表 3 高频关键词的绝对词频与相对词频差异统计

关键词	绝对词频						相对词频					
	2016 年	2015 年	2014 年	2013 年	2012 年	累计总和	2016 年	2015 年	2014 年	2013 年	2012 年	累计总和
数据库	9	13	17	25	37	101	0.000	0.001	0.002	0.007	0.006	0.018
情报分析	23	20	20	17	21	101	0.006	0.004	0.004	0.002	0.006	0.023
信息需求	18	13	20	22	25	98	0.003	0.001	0.004	0.005	0.006	0.020
社交网络	23	18	18	22	18	99	0.007	0.003	0.003	0.005	0.006	0.024
读者服务	17	15	16	29	30	107	0.002	0.002	0.002	0.011	0.006	0.023
微信	39	28	31	11	0	109	0.029	0.011	0.014	0.001	0	0.055
移动服务	19	8	28	26	11	92	0.004	0.000	0.012	0.009	0.006	0.032
用户需求	14	13	25	19	20	91	0.002	0.001	0.009	0.004	0.006	0.022
信息资源	14	11	22	44	48	139	0.001	0.001	0.004	0.029	0.037	0.072
信息行为	22	28	36	30	23	139	0.004	0.009	0.017	0.009	0.004	0.043

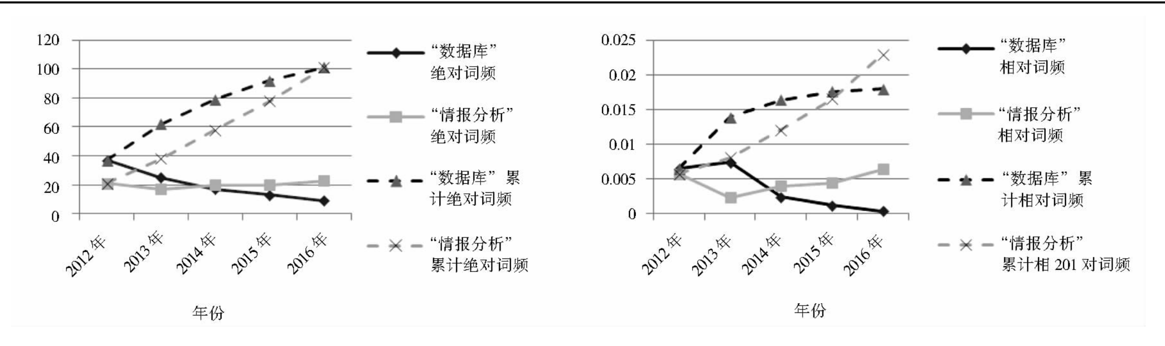


图 3 2012 - 2016 年“数据库”“情报分析”的绝对词频及累计量（左）和相对词频及累计量（右）

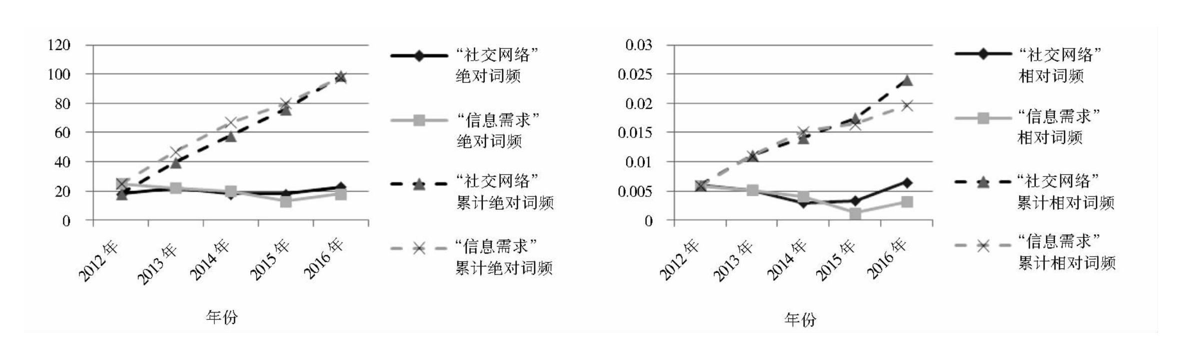


图4 2012－2016年"社交网络""信息需求"的绝对词频及累计量(左)和相对词频及累计量(右)

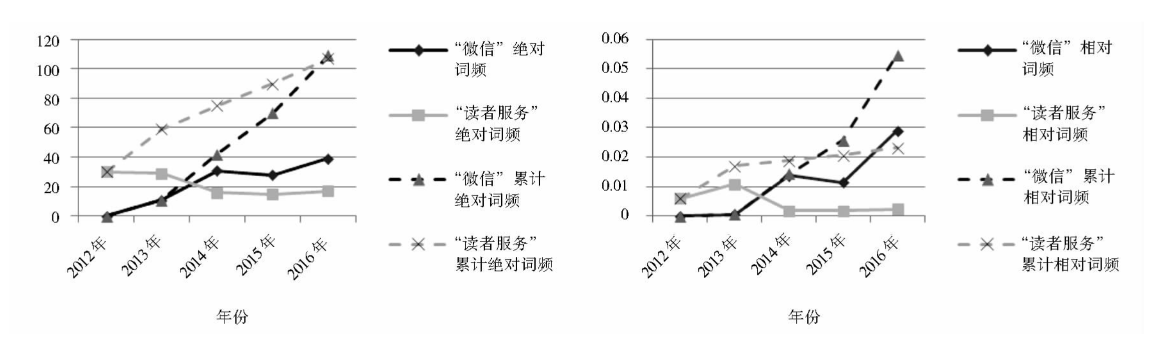


图5 2012－2016年"微信""读者服务"的绝对词频及累计量(左)和相对词频及累计量(右)

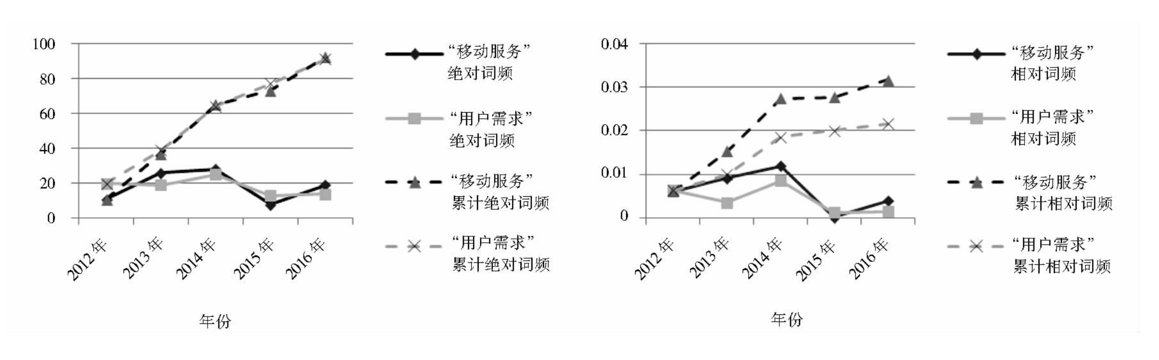


图6 2012－2016年"移动服务""用户需求"的绝对词频及累计量(左)和相对词频及累计量(右)

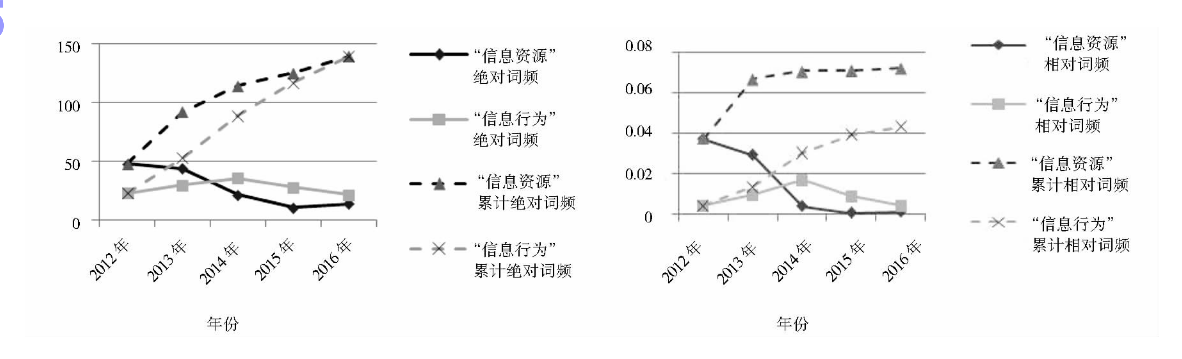


图7 2012－2016年"信息资源""信息行为"的绝对词频及累计量(左)和相对词频及累计量(右)

力,因此"微信"累计相对总词频较"读者服务"更为突出(见图5)。④"移动服务"和"用户需求",两者均为波动型关键词,前者峰值高于后者,突出"移动服务"的总体优势,且累计相对总词频较大(见图6)。⑤"信息资源"和"信息行为",前者峰值优势明显,虽2015、2016年度数值不高,其排名仍上浮(见图7)。

针对高频词,加权关键词相对词频排序中,上浮关键词是"量高质优"型和"量中质优"型,也即增长趋势大、峰值优势明显、高频且稳定的关键词。基于Logistic增长规律,概念频次大幅度增长为新兴概念,文献频次增速渐缓则概念达到成熟期[13]。因此,利用加权关键词相对词频模型,可以快速并客观地找出品质好

的关键词,挖掘出具有发展潜力的关键词,进而揭示学科热点和预测发展趋势。

3.3.2　低频关键词实验结果　主题突变是指在某一领域中,随着某一事件的发生在短时间内引起关注度改变的主题变化情况。随着时间推移,突变主题有可能变成研究热点,也有可能趋弱为普通主题甚至消逝。因此,对突变词的监测是有重要意义的。根据突变词出现频次的时间变化,将主题突变类型分为上升型、下降型、先升后降型、突现型、稳定型[14]。而低频关键词的排名变化与关键词突变之间有密切关系。

分别在每个排名变化量阶段选取部分关键词,对比低频词前后排名变化情况(见表 4 和图 8)。其中,负数为该关键词在相对词频中排名低于绝对词频排名,正数则相反。

表 4　低频关键词的绝对词频与相对词频差异统计

关键词	排序差	绝对词频					相对词频				
		2016 年	2015 年	2014 年	2013 年	2012 年	2016 年	2015 年	2014 年	2013 年	2012 年
数据素养教育	-1064	12	0	0	0	0	0.000 12				
移动社交网络	-906	8	2	0	0	0	0.000 05	0.000 05			
文化扶贫	-856	8	0	0	0	0	0.000 05				
文本相似度	-461	4	0	0	3	1	0.000 07			0.000 06	0.000 01
libraries	-328	0	20	0	6	0		0.000 27		0.000 40	
儿童图书馆	-289	0	4	4	5	0		0.000 12	0.000 10	0.000 10	
政务信息资源	-247	8	0	0	3	0	0.000 13			0.000 12	
潜在语义分析	100	2	2	3	0	0	0.000 06	0.000 06	0.000 06		
网络计量学	112	3	4	3	2	5	0.000 21	0.000 23	0.000 20	0.000 19	0.000 19

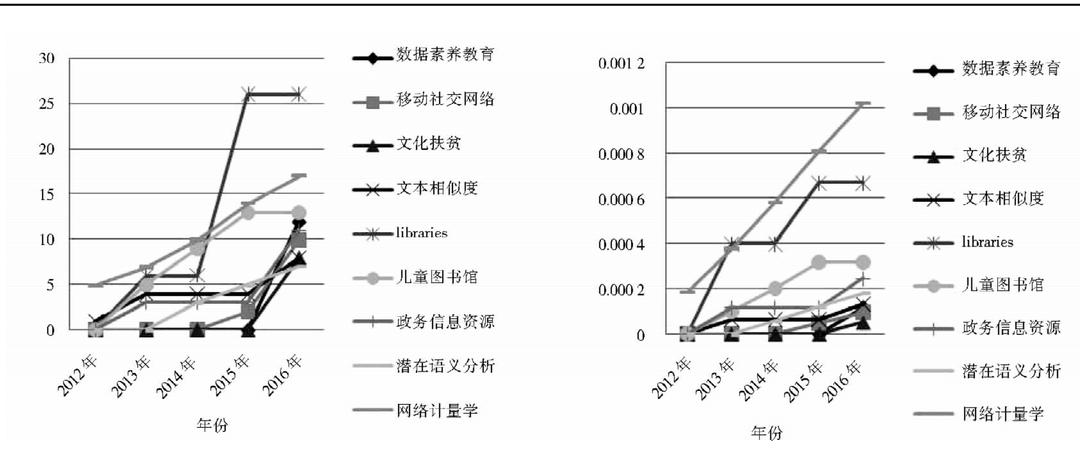


图 8　2012 -2016 五年间部分低频关键词累计绝对词频折线图(左)和累计相对词频折线图(右)

结果表明,排名变化突出的关键词多表现为骤升骤降,由于低频词的词频变化比高频词更为敏感,骤升骤降的频次导致词频变化显著,而关键词排名也产生显著变化。因此,通过统计低频关键词排名变化量可以侧面探测突变词,以及总结出排名变化量与突变主题类型关系的表征关系。从表 4 可以看出,“数据素养教育”“libraries”“政务信息资源”累计绝对词频量低,然而近 5 年出现突增(累计相对词频量高),有望成为未来的研究趋势。

排名下降量显著的关键词主要呈现为突显型主题突变。如数据素养教育、移动社交网络、文化扶贫,骤升骤降的频次导致突变机会较大,这表明此类研究有着社会时效性,同时研究热度逐渐上升。

排名上升量显著的关键词主要呈现为稳定型主题突变,如潜在语义分析、网络计量学,这表明其频次波动不大或者频次突增,也表明该研究逐渐稳定,未来也处于稳定的发展状态中,或者该研究视角已经结合其他学科内容成为新的研究主题。

4　结论

针对目前基于对词频的简单统计和粗略分析,以揭示学科领域热点及趋势的普遍情况,笔者提出加权关键词相对词频 WRKFM 模型,构建年度 - 关键词的词频矩阵,依据矩阵水平与垂直两个维度加权处理关键词词频,导出相对词频计算公式,得到关键词加权综合分值,以获得更有效的关键词排序。从而更为准确、

客观、科学地揭示学科热点和趋势，并通过其数据结果揭示出学科领域知识的内在规律。实证表明，笔者提出的 WRKFM 模型，相较于绝对词频分析法，具有以下特征：①具有“量高质优”和“量中质优”的高频词排名靠前，其具有发展前景，可以揭示学科热点；②骤升骤降的低频词排名较增幅或降幅显著，利用排名变化量侧面探测突变词，进而预测研究热点与趋势，有利于挖掘突变型关键词。

本文初步实现了预期的设计目标，但还存在需要进一步研究和修改的地方，主要包括以下几点：①该实验仅选取文献给出的关键词，下一步数据采集可扩充到标题关键词、摘要关键词、全文关键词，以期提高结果的准确性和全面性；②WRKFM 模型计算排序分数受该关键词时域词频峰值的影响，若某年该关键词词频突增，则分数较高，而对稳定型关键词不利；③该模型中只对年度－关键词分布及占比进行计算分析，并未涉及词义的加权，也未实现“核心关键词上浮、辅助性关键词下浮”，之后需要通过借助其他方法来做进一步的加权判断，实现自动识别核心关键词。

参考文献：

[1] 李文兰，杨祖国. 中国情报学期刊论文关键词词频分析[J]. 情报科学,2005(1):68－70,143.
[2] 张勤. 词频分析法在学科发展动态研究中的应用综述[J]. 图书情报知识,2011(2):95－98,128.
[3] 田丹，刘奕杉，王玉琳. 热点分析类文章的文献计量分析——以词频分析方法为例[J]. 情报科学,2017,35(8):164－169.
[4] 巩永强，刘莉. 基于词频分析法的情报学研究热点透析[J]. 图书馆学研究,2011(13):9－13.
[5] 安兴茹. 我国词频分析法的方法论研究(Ⅰ)——统计分析要素的界定、分类及问题[J]. 情报杂志,2016,35(2):75－80,43.
[6] 倪丽娟，于淑丽. 档案学研究热点分析——基于 2004－2008 年《档案学研究》、《档案学通讯》论文关键词的词频分析[J]. 档案学通讯,2010(1):19－22.
[7] 邱均平，丁敬达. 1999－2008 年我国图书馆学研究的实证分析(下)[J]. 中国图书馆学报,2009,35(6):79－87,118.
[8] 苍宏宇，谭宗颖. 国内外信息检索研究热点分析——基于 Z-Score 标准化的词频[J]. 图书馆建设,2009(1):93－98.
[9] CHEN G, XIAO L, ZHAO X G. A keyword selection method based on the combination of popularity and domain relevancy of keywords: a holistic perspective[J]. Journal of the China Society for Scientific and Technical Information,2014,33(9): 959－968.
[10] ATLAM E S, FUKETA M, AOE J, et al. Similarity measurement using term negative weight and its application to word similarity[J]. Information processing & management,2000,36(5): 717－736.
[11] CHEN G, XIAO L. Selecting publication keywords for domain analysis in bibliometrics: a comparison of three methods[J]. Journal of informetrics,2016,10(1): 212－223.
[12] 李姗姗，张国强，徐桂芬. 基于关键词分析的 ERP 系统研究热点评述[J]. 情报科学,2012,30(8):1272－1276.
[13] 杨婧，常春. 基于 Logistic 种群增长规律的概念词频变化研究[J]. 情报科学,2017,35(8):15－18,50.
[14] 王梦婷. 基于突变检测的主题突变分析研究[J]. 情报科学,2016,34(12):36－39.

作者贡献说明：

奉国和：提出研究思路，设计研究方案；
孔泳欣：进行实验，采集、清洗和分析数据，起草论文；
肖洁琼：负责最终版本修订。

A New Model for Hotspot and Trend Analysis Based on Weighted Keywords

Feng Guohe[1] Kong Yongxin[2] Xiao Jieqiong[1]
[1] The Department of Information Management, School of Economics & Management, South China Normal University, Guangzhou 510006
[2] Department of Information Resources Management, Business School, Nankai University, Tianjin 300071

Abstract: [**Purpose/significance**] In order to overcome the limitation of the absolute word frequency analysis of the keywords, the hot spots and trends in the field are explored by using the multi-factor weighting and ranking of the keywords. [**Method/process**] It constructs the annual-key frequency matrix, processes the word frequency of horizontal and vertical weighting, and derives the formula of relative word frequency to get the weighted comprehensive score of keywords, in order to obtain more effective keyword ranking. [**Result/conclusion**] Based on keyword weighted ranking, three types of keywords, including keywords in large quantities & high quality, keywords in little quantities & high quality, and burst terms, can be identified, greatly benefiting focus mining and trends analysis.

Keywords: keyword frequency analysis weighted keywords hotspot research trend analysis